On the Exact Distribution and Moment-Coefficients of the D^2-Statistics

Given a number of normal populations it is often desirable to have a numerical measure of the divergence between two samples drawn from them. For example, if we have three samples Σ_1, Σ_2, Σ_3, we should like to know whether the difference between Σ_1 and Σ_2 is significantly greater than the difference between Σ_1 and Σ_3. Professor P. C. Mahalanobis has given a generalised statistical measure of such divergence between two P-variate samples. This statistics, which he calls D^2, has proved extremely useful in biometric studies[1]. He obtained the first four moments of D^2 by approximate methods for the case of uncorrelated variates[2]. The exact distribution of D^2 and a general expression for the higher moments were not however known. I have now succeeded in obtaining these results which are given below. Fuller details will be published shortly in *Sankhya: The Indian Journal of Statistics.*

If Δ^2 denotes the population value of D^2, and D_1^2 the uncorrected sample value of the same statistic then the kth moment of D_1^2 is given by

$$\mu'_k(D_1^2)=\Delta^{2k} \underset{c\to\infty}{Lt}\ F\left(-k, -k-\frac{P}{2}+1, c, \frac{4c}{nP\Delta^2}\right) \quad \ldots\ldots\ldots(1)$$

where F is the well-known hypergeometric function, and n is the harmonic mean between n_1 and n_2, the number of individuals in the two samples. Since

$$\mu_1'(D_1^2) = \Delta^2 + \frac{2}{n} \quad \ldots\ldots\ldots(2)$$

it is clear that to get rid of the bias in the mean we must put

$$D^2 = D_1^2 - \frac{2}{n} \quad \ldots\ldots\ldots(3)$$

The first four moments of D^2 turn out to be exactly those given by Mahalanobis[2] showing that (*i*) his expressions are exact, in spite of the fact that they were obtained by approximate methods, and that (*ii*) his expressions remain valid for the case of correlated variables.

The exact distribution of D_1^2 comes out to be

$$f(D_1^2)\ d(D_1^2) = \frac{nP}{4}\left(\frac{D_1^2}{\Delta^2}\right)^{\frac{P-2}{4}} e^{-\frac{nP}{4}(D_1^2+\Delta^2)}$$

$$\times\ I_{\frac{P-2}{2}}\left(\frac{nPD_1\Delta}{2}\right) d(D_1^2) \ldots\ldots\ldots(4)$$

where $I_n(x)$ is the well-known Bessel function with purely imaginary argument. The distribution of D^2 is obtained by substitution from (*3*).

In the special case when $\Delta^2=0$, the above distribution reduces to

$$\frac{1}{\Gamma\left(\frac{P}{2}\right)}\left(\frac{Pn}{4}\right)^{\frac{P}{2}}(D_1^2)^{\frac{P-2}{2}} e^{-\frac{nPD_1^2}{4}} d(D_1^2) \quad \ldots\ldots(5)$$

If we put $m=\frac{P-2}{2}$, $x=\frac{nPD_1^2}{4}$, then the distribution of x can be written as:

$$\frac{1}{\Gamma(m+1)} x^m\ e^{-x}\ dx \quad \ldots\ldots\ldots(6)$$

The probability of attaining any assigned value of x (when $\Delta^2=0$) can be obtained immediately from a table of incomplete Gamma functions. When however Δ^2 is not zero, the probability of D_1^2 attaining any given value is given by

$$\int_0^{D_1^2} f(D_1^2)\, d(D_1^2) \quad \ldots\ldots\ldots(7)$$

where $f(D_1^2)$ is given in (*4*).

A table of the values of the above integral will be of great practical importance in many branches of applied statistics. We are compiling such a table by rather tedius computations in the Statistical Laboratory. I should like to draw the attention of pure mathematicians to the evaluation of this integral. A closed expression would not perhaps be forthcoming. If it does, so much the better. If not, a convergent series (preferably a rather rapidly convergent one) would do. But the same series may not be useful for computational purposes for all ranges of values of the independent variable and the parameters involved. In that case different series for different regions (confining ourselves, of course, to those regions which are of practical interest) would have to be used.

Statistical Laboratory,
Presidency College,
Calcutta. 7.7.35.

Rajchandra Bose.

1. Analysis of Race-Mixture in Bengal. *Ind. Science Congress* 1926, and *Jour. Asiat. Soc. Bengal*, 23 1927.
Statistical Study of Chinese Head. *Man in India*, 8, 1928.
2. Tests and Measures of Divergence. *J.A.S.B.* 26, 1930, 4.